COURSE : Opening the Black Box of Generative AI
SEMESTER/YEAR.... 2º/2026
PROGRAM..... Winter School
CLASS-HOURS..... ☑ 30 hours
PROFESSOR : Erwan LAMY
LANGUAGE.....ENGLISH

COURSE OBJECTIVES

Generative AI technologies are often perceived as "black boxes" due to their complex, opaque internal workings. The aim of this course is to begin to open up this black box, to gain a better understanding of how this technology works. This course provides an in-depth exploration of generative AI, emphasizing the detailed understanding of neural networks and their role in generating diverse outputs such as text, images, and sound. Students will critically examine the mechanisms underpinning generative models, from basic neural structures to advanced architectures like GPT. The course includes theoretical lectures, providing essential concepts and analytical frameworks to deeply understand generative models, and practical workshops, offering hands-on experiences and collaborative exercises to build and interpret these models concretely. The course also thoroughly addresses significant intellectual and ethical implications, exploring the limitations, biases, and ethical responsibilities involved in using and interpreting generative AI. Students will develop informed perspectives and responsible strategies for the ethical and intellectual employment of generative AI technologies.

LEARNING GOALS

The course learning goals are presented in the table below, showing how they contribute to the learning goals related to methods, for the stricto sensu graduate programs at FGV-EAESP.

Program learning goals	**Course learning goals**	**Level of contribution**
Scientific method	---	○ ○ ○
Research project / procedures	Grasping the workings and limitations of generative AI technologies Developing responsible usage strategies for generative AI Fostering ethical awareness and responsible use of generative AI Developing intellectual integrity in interacting with AI technologies Being able to critically assess the intellectual capabilities and limitations of generative AI	● ● ○
Qualitative research methods	---	○ ○ ○
Quantitative research methods	---	○ ○ ○
Development of academic papers	To be able to use generative AIs ethically and effectively for scientific writing practices	● ● ●
<u>Other course learning goals:</u> Understanding neural network architectures and generative models Being able to construct and manually simulate simple neural networks Being able to identify and explain key concepts like transformers, embeddings, attention mechanisms, and RLHF		

The full description of the learning goals of FGV-EAESP stricto sensu graduate programs can be found at https://rebrand.ly/cursos-pos-eaesp.

PREVIOUS KNOWLEDGE REQUIRED

None

CONTENT/METHODOLOGY

Session 1: AI and neural networks

With this first session, I start by laying the conceptual foundations for understanding the notions of AI and generative AI, and for situating generative AI within the broader field of AI. Then I present the technological and mathematical foundations common to all generative AI systems.

- Historical background
  From the precursors of past centuries to ChatGPT, via Turing. With this historical overview, I will highlight the diversity of AIs and situate generative AIs within it
- The two main paradigms of AI
  There are two main paradigms in AI: the symbolic paradigm, which involves formalizing human reasoning, and the connectionist paradigm, which consists in imitating the neural structure of the human brain.
- The connectionist paradigm
  I will explain what this paradigm entails, what an artificial neuron is, including its mathematical basis (requiring only a high school level of math).
- What is a neural network?
  Generative AIs are artificial neural networks. I will present several examples of such networks and explain how they work.
- Loss functions and training methods
  Once the structure of a neural network is defined (through programming), these systems are not programmed in the traditional sense, but trained. I will explain what training means and the various existing training methods. To understand how these networks are trained, I will also introduce the concept of a loss function. I will also explain the notions of deep learning and machine learning.

Session 2: Workshop

The theoretical teaching of neural network basics is generally not sufficient to fully understand this technology: hands-on experience is essential to grasp how it really works.

- Group-based creation of neural networks using Python and Excel
  In groups of three or four, students will build two neural networks. The first, extremely simple, should be able to locate points on a divided plane. The second, slightly more advanced, should be able to count the number of "lit" cells in a table. These neural networks will be trained using Python and implemented in Excel.
- Presentation and discussion of created models

Session 3: Image and sound generative models (Introduction and key concepts)

The most important generative AIs in business contexts are those that produce text, for obvious reasons. However, one should not overlook other types of generative AIs that produce images, videos, or sound. Without going into as much detail as for text-generating AIs, I will present their underlying technologies, the different existing types, and their economic, ethical and political implication.

Session 4: Text-generating models (GPT kind) part I

Text-generating AIs like GPT are autoregressive language models, pre-trained on large textual corpora, based on the Transformer architecture, and optimized for natural language generation. (The term "GPT" should be reserved for OpenAI's specific models, but other major generative AIs such as Claude, Gemini, LLaMA, etc., share the same architectural type.) These AIs rely on very specific technical concepts that must be understood in order to use them properly. These are the concepts I will present in this session and the following two.

- Introduction to LLMs
  The notion of Large Language Models (LLMs) is foundational to text-generating AI. In order to produce text, these machines must "model" human language. I will explain what this means.
- Transformers
  The Transformer architecture, invented in 2017, enabled the emergence of generative AIs known to the

general public since 2022. I will start by providing a general presentation, before detailing its three main components: embedding, attention, and classifier.

- Embeddings
  The embedding technology used in Transformers was, even before their invention (into which it is now integrated), a technological revolution in natural language processing. I will explain its nature.

Session 5: Text-generating models (GPT kind) part II

- Attention mechanisms
  The attention mechanism is at the heart of Transformer technology. It is what makes this architecture so original and powerful. Very schematically, it allows the system to account for the fact that the meaning of a word or sentence depends on its context.
- Classifier
  At the end of the processing chain, Transformers calculate the probability that a word can complete a text. I will explain how this calculation is made.
- Autoregression
  Text-generating AIs are autoregressive LLMs, which means they reuse their own output as input and iteratively complete it until the text is finalized. I will explain and illustrate this concept of autoregression.

Session 6: Text-generating models (GPT kind) part III

- Alignment
  The concept of alignment is ethical in nature, but it is nonetheless crucial from a technical point of view. Generative AIs are extremely powerful and potentially very dangerous. It is therefore essential to ensure they are "aligned" with human interests and values. This is not only a philosophical issue, but also a particularly difficult technical challenge.
- Training and RLHF
  Like any neural network, text-generating AIs must be trained, but in a manner specific to them, so as to guarantee their alignment. In particular, these AIs undergo a specific training called "Reinforcement Learning through Human Feedback" (RLHF), which helps correct undesirable biases.
- Exercise: How does ChatGPT create anagrams?
  At the end of all these explanations, I will ask students to explain how ChatGPT can generate an anagram. This is indeed a difficult question, which requires understanding GPT-type technology (and, incidentally, is one to which AIs respond very poorly).

Session 7: Epistemic limitations of GPT and ethical implications

Understanding how text-generating AIs function makes it possible to grasp their limitations and therefore to use them appropriately. In this session, I will revisit these limitations and explain how they are directly linked to the current architecture of these systems.

- Review and discussion of previous session exercise
  I will revisit the mistakes made by students in the exercise given at the end of Session 7 and explain how text-generating AIs can probably compute an anagram, based on previously discussed content.
- Bullshit engines
  Building on the previous explanations of how text-generating AIs function, I will explain why they are "bullshit engines," why they are still often reliable, and what this implies for how we use them. I will explain the philosophical notion of bullshit.
- Issues with hallucinations and cognitive biases
  The fact that text-generating AIs are bullshit engines manifests concretely through hallucinations. I will explain what these hallucinations are, how they differ from errors and biases, and revisit the very important issue of biases.
- Normativity problem in AI
  These concrete technological problems are difficult to solve because they originate from a highly complex philosophical question: the naturalization of epistemic normativity. I will explain this problem, outline its complexity (including historically), and consider a few possible solutions.
- Intellectual ethics of generative AI (Guidelines for responsible intellectual use of AI)
  In conclusion of this course, I will propose an intellectual ethics of generative AI — that is, a way to rationally determine whether the use of these tools to produce knowledge is acceptable, and to define what constitutes appropriate use of these systems.

Session 8: Writing workshop
In groups of three or four, students write a short scientific paper on a topic of their choice, using the generative AIs at their disposal and building on the knowledge acquired in previous sessions. In a second stage of the session, each group hands the result of their work to another group, which will read the paper and prepare a critical analysis for the final part of the session. During this final part, one or two papers are presented, followed by a collective debrief and discussion of the scientific and ethical issues raised by this writing practice.

EVALUATION CRITERIA

Assessment component	Assessment type	Weighted %	Pass Mark %	Pass/Fail
Manual neural network creation project	Group work	20%	-	No
Anagram exercise and presentation	Group work	30%	-	No
Final exam	Multiple choice questions exam	50%	40%	Yes

COURSE SCHEDULE (OPTIONAL)

BIBLIOGRAPHY

- Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. Ethics and Information Technology, 26(2), 1-10.
- Liang, K., Hu, H., Zhao, X., Song, D., Griffiths, T. L., & Fisac, J. F. (2025). Machine Bullshit: Characterizing the Emergent Disregard for Truth in Large Language Models. arXiv preprint arXiv:2507.07484.
- Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier models are capable of in-context scheming. arXiv preprint arXiv:2412.04984.
- Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, *5*(5).
- Wolfram (2023), What Is ChatGPT Doing… and Why Does It Work? https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

FURTHER INFORMATION (OPTIONAL)